1. Pendahuluan

Dalam bidang bioinformatik, *Multiple Sequence Alignment* (MSA) adalah salah satu metode untuk menyusun rantai protein menjadi satu konsesus rantai protein baru yang memiliki panjang *sequence* sama. MSA memiliki kemampuan untuk menyusun rantai protein lebih dari satu pasang. Peranan penting dari MSA dalam bidang bioinformatik digunakan untuk menghasilkan informasi biologis yang lebih banyak dibandingkan dengan *pairwaise alignment*.[1] Hal tersebut menjadikan MSA sebagai suatu proses yang penting dilakukan didalam dunia bioinformatik.

Saat ini telah terdapat beberapa aplikasi yang dapat kita gunakan untuk menyusun rantai protein menggunakan metode MSA. Aplikasi tersebut antara lain adalah PRRN yang menggunakan pendekatan *iterative alignment*, DIALIGN2 yang menggunakan pendekatan *block-based aligment*, dan clustal yang menggunakan pendekatan *progressive alignment*. Dari beberapa aplikasi yang dapat melakukan MSA diatas, clustal adalah aplikasi yang lebih sering digunakan. Penyebabnya adalah karena clustal dapat dijalankan secara multi *platform*, baik didalam OS windows maupun OS Linux. Selain hal tesebut, yang menjadikan clustal lebih banyak digunakan karena clustal memiliki versi *graphic* yang lebih ramah terhadap pengguna.

Walau cukup popular, clustal memiliki kekurangan yaitu pada lama proses. Ketika jumlah rantai protein yang akan digunakan lebih dari 100 rantai protein, proses komputasi yang dilakukan oleh clustal menjadi cukup lambat.[2]. Fakta tersebut diakibatkan karena clustal harus melakukan tiap tahapan dari MSA dengan pendekatan *progressive alignment* secara sekuensial. Hal diatas tentu menjadi masalah yang serius di masa yang akan datang karena jumlah rantai protein yang dapat diproses oleh clustal akan meningkat secara pesat di masa depan. [2]

Hadoop sebagai lingkungan pengolahan data besar memiliki kemampuan untuk melakukan proses komputasi dengan jumlah data yang sangat besar menjadi lebih cepat.[5] Hal tersebut diakibatkan karena hadoop memisahkan tugas komputasi dan tugas penyimpanan data kedalam dua *server* yang berbeda. Tugas komputasi dikerjakan oleh mapreduce sedangkan tugas penyimpanan data dilakukan oleh HDFS. Didalam mapreduce, tugas besar yang diberikan kepada *server* kemudian akan dipecah kembali menjadi tugas – tugas kecil yang akan dikerjakan oleh *client*. Begitu juga sebaliknya yang dilakukan oleh HDFS, sebuah data yang berukuran sangat besar akan dipecah – pecah menjadi potongan kecil agar dapat disimpan didalam *client*.

Dengan keunggulan yang dimiliki oleh hadoop maka permasalahan pada clustal yang cukup memakan waktu ketika harus memproses rantai protein yang berjumlah banyak dapat diatasi dengan model komputasi multiple sequence alignment yang dapat diterapkan didalam lingkungan hadoop.

.

Topik dan Batasannya

Berdasarkan penjelasan pada latar belakang, maka dirumuskan masalah yang dibahas dalam penelitian model komputasi multiple sequence alignment dalam lingkungan hadoop seperti berikut :

1) Bagaimana implementasi clustal W di dalam lingkungan Hadoop ?
2) Bagaimana performa yang dihasilkan sistem ?
3) Bagaimana performa clustalW di dalam lingkungan Hadoop dibandingkan performa clustalW tanpa menggunakan Hadoop ?

Adapun batasan masalah dalam melakukan penelitian ini adalah :

1) Menggunakan clustalW,
2) Menggunakan *Hadoop distribution Cloudera*,
3) Menggunakan arsitektur Mapreduce V.1.

Tujuan

Yang menjadi tujuan dari pengerjaan tugas akhir berikut adalah untuk mencapai tiga tujuan utama. Pertama adalah untuk memaparkan implementasi model komputasi *multiple sequence alignment* di dalam lingkungan Hadoop. Kedua adalah untuk menjelaskan performa yang dihasilkan oleh sistem. Ketiga adalah untuk menjelaskan perbandingan performa clustalW di dalam lingkungan Hadoop dibandingkan dengan performa clustalW tanpa menggunakan Hadoop.